TBox Reasoning in the Probabilistic Description Logic $\mathcal{SHIQ}_\mathcal{P}$

Viachaslau Sazonau and Uli Sattler

The University of Manchester
Oxford Road, Manchester, M13 9PL, UK
{sazonauv, sattler}@cs.manchester.ac.uk

Abstract. One shortcoming of classic Descriptions Logics, DLs, is their inability to encode probabilistic knowledge and reason over it. This is, however, a strong demand of some modern applications, e.g. in biology and healthcare. Therefore, probabilistic extensions of DLs are attracting attention nowadays. We introduce the probabilistic DL $\mathcal{SHIQ}_\mathcal{P}$ which extends a known probabilistic DL. We investigate two reasoning problems for TBoxes: deciding consistency and computing tight probability bounds. It turns out that both problems are not harder than reasoning in the classic counterpart $\mathcal{SHIQ}$. We gain insight into complexity sources.

1 Introduction

Descriptions Logics [1], DLs, are a family of knowledge representation formalisms that form the basis for popular knowledge representation languages. In particular, they underly the Web Ontology Language [5], OWL, which is a W3C standard. Logical theories that encode a domain of interest in such languages are called ontologies. The last decade has witnessed a rapid growth in the number and size of ontologies which have become the common way to encode and share information in application areas such as medicine, biology, astronomy, defence and others.[1] Since DLs are essentially decidable fragments of first-order logic, FOL, ontologies are only capable to encode certain knowledge. Although some means of uncertainty, in fact, can be encoded, e.g. "a parent is a human who has some children" and "sky is sunny or cloudy", there are no built-in ways to represent probabilistic knowledge.

Successful application areas of DLs, however, often require modelling probabilistic knowledge which DLs are not able to deal with. The examples showing this appeal are evident in medicine. For instance, the medical ontology SNOMED CT [17] contains concepts mentioning "Probable cause", "Probable diagnosis", etc. This shortcoming in expressive power of classic DLs has recently caused non-classic proposals and various extensions. A recent survey is given in [12].

Two main approaches to probabilistic extensions of DLs differ in the view of probability. Halpern [4] makes a distinction between *statistical* and *subjective* probabilities. He formalizes statistical probabilities in "Type 1" probabilistic FOL and subjective probabilities in "Type 2" probabilistic FOL. The statistical view considers a probability

[1] http://bioportal.bioontology.org/

distribution over a *domain* that specifies the probability for an individual in the domain to be randomly picked. The subjective view is based on so-called *possible worlds* and specifies the probability distribution over a set of possible worlds [4]. We call the semantics of a probabilistic DL *statistical* if it uses the statistical view (Type 1 extension) and *subjective* if it is based on possible worlds (Type 2 extension).

The contributions of this work are as follows. Firstly, we introduce a new probabilistic extension of the classic DL $\mathcal{SHIQ}$, called $\mathcal{SHIQ}_\mathcal{P}$, with the statistical semantics inherited from the Type 1 probabilistic FOL that distinguishes it from many existing extensions whose semantics is based on possible worlds. We discuss relations to other extensions and some important features of the statistical semantics of $\mathcal{SHIQ}_\mathcal{P}$. Secondly, we study two reasoning problems for TBoxes in this extension: deciding consistency and computing tight probability bounds. We show that both problems are in ExpTime in the size of a TBox which implies that they are not harder than reasoning in the classic DL $\mathcal{SHIQ}$. The algorithm for solving each problem consists of two parts: detecting satisfiable types and solving a linear program on those types. We show that, in fact, the linear program can be built on the types over the probabilistic part only. Some examples and proofs are moved to Appendix.[2]

2 Syntax and Semantics of $\mathcal{SHIQ}_\mathcal{P}$

The syntax of $\mathcal{SHIQ}_\mathcal{P}$ extends classical $\mathcal{SHIQ}$ axioms with probabilistic statements over concepts. We assume the reader to be familiar with the DL $\mathcal{SHIQ}$ [6]. As usual, a classic $\mathcal{SHIQ}$ TBox $\mathcal{T}_c$ is a finite set of general concept inclusions and role inclusions.

Definition 1. *(TBox syntax) A $\mathcal{SHIQ}_\mathcal{P}$ TBox $\mathcal{T}$ is a set $\mathcal{T}_c \cup \mathcal{T}_p$, where $\mathcal{T}_c$ is a classic $\mathcal{SHIQ}$ TBox and $\mathcal{T}_p$ is a set of probabilistic statements. A probabilistic statement is a statement in one of the following forms:*

(i) $\sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) \bowtie r'$ *(unconditional form);*

(ii) $\sum_{j=1}^{m''} a''_j \cdot \mathbb{P}(C_j|D) \bowtie r''$ *(conditional form);*

where $\bowtie\ \in \{<, \leq, \geq, >\}$, $a'_j, a''_j, r', r'' \in \mathbf{R}$, $C_j,\ D$ *are possibly complex $\mathcal{SHIQ}$ concepts. We call concept inclusions, role inclusions, and probabilistic statements* axioms *of a TBox.*

We use the abbreviation $C \equiv D$ for $\{C \sqsubseteq D,\ D \sqsubseteq C\}$ and $\sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) = r'$ for $\{\sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) \leq r',\ \sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) \geq r'\}$ in (i) and the analogous one in (ii). Before defining the semantics we give an illustrative example, see Example 1.

Example 1. ($\mathcal{SHIQ}_\mathcal{P}$ TBox) According to statistics on smoking in England in 2010,[3] 47% of all adults in the sample are men (1); 20% and 25% of all adults (2), 20% and 29% of all men (3), 19% and 22% of all women (4) are current and former smokers, respectively. Current and former smokers are 17% and 28% of all married adults (5),

[2] www.cs.man.ac.uk/~sazonauv/SazonauDL15.pdf

[3] http://www.hscic.gov.uk/catalogue/PUB11454

18% and 33% of all married men (6), 17% and 23% of all married women (7). 34% and 25% of all diseases are caused by smoking for men and women (8), respectively. This knowledge can be expressed as follows:

$$\begin{aligned}
&\mathcal{T} = \{S \sqsubseteq A,\ CS \sqsubseteq S,\ FS \sqsubseteq S,\\
&FS \sqsubseteq \neg CS,\ S \sqsubseteq CS \sqcup FS,\\
&M \sqsubseteq A,\ W \sqsubseteq A,\ M \sqsubseteq \neg W,\ A \sqsubseteq M \sqcup W,\\
&\mathbb{P}(M \mid A) = 0.47, && (1)\\
&\mathbb{P}(CS \mid A) = 0.2,\ \mathbb{P}(FS \mid A) = 0.25, && (2)\\
&\mathbb{P}(CS \mid M) = 0.2,\ \mathbb{P}(FS \mid M) = 0.29, && (3)\\
&\mathbb{P}(CS \mid W) = 0.19,\ \mathbb{P}(FS \mid W) = 0.22, && (4)\\
&\mathbb{P}(CS \mid \exists m.A) = 0.17, \mathbb{P}(FS \mid \exists m.A) = 0.28, && (5)\\
&\mathbb{P}(CS \mid M \sqcap \exists m.W) = 0.18,\ \mathbb{P}(FS \mid M \sqcap \exists m.W) = 0.33, && (6)\\
&\mathbb{P}(CS \mid W \sqcap \exists m.M) = 0.17,\ \mathbb{P}(FS \mid W \sqcap \exists m.M) = 0.23, && (7)\\
&\mathbb{P}(\exists d.(D \sqcap \exists c.S) \mid M) = 0.34,\ \mathbb{P}(\exists d.(D \sqcap \exists c.S) \mid W) = 0.25\}, && (8)
\end{aligned}$$

where A, M, W, S, CS, FS, D are concepts representing adults, men, women, smokers, current smokers, former smokers, diseases, respectively; m, d, c are roles representing marriage, having a disease, having a cause, respectively.

The semantics of a $\mathcal{SHIQ}_\mathcal{P}$ TBox is based on the statistical view of probability and inherited from Halpern's Type 1 probabilistic FOL [4].

Definition 2. *(TBox semantics) An* interpretation *of a $\mathcal{SHIQ}_\mathcal{P}$ TBox is a structure $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I}, \mu)$ where $(\Delta^\mathcal{I}, \cdot^\mathcal{I})$ is a standard $\mathcal{SHIQ}$ interpretation and μ a discrete*[4] *probability distribution over $\Delta^\mathcal{I}$. The semantics of concept and role inclusions is defined as usual. For concepts C, D we set $P(C^\mathcal{I}) = \sum_{d \in C^\mathcal{I}} \mu(d)$ and*

$$P(C^\mathcal{I}|D^\mathcal{I}) = \begin{cases} \frac{P(C^\mathcal{I} \cap D^\mathcal{I})}{P(D^\mathcal{I})} & \textit{if } P(D^\mathcal{I}) > 0 \\ 0 & \textit{otherwise} \end{cases}$$

An interpretation $\mathcal{I}$ satisfies a probabilistic statement

(i) $\sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) \bowtie r'$ *if* $\sum_{j=1}^{m'} a'_j \cdot P(C_j^\mathcal{I}) \bowtie r'$ *holds;*

(ii) $\sum_{j=1}^{m''} a''_j \cdot \mathbb{P}(C_j|D) \bowtie r''$ *if* $\sum_{j=1}^{m''} a''_j \cdot P(C_j^\mathcal{I}|D^\mathcal{I}) \bowtie r''$ *holds.*

An interpretation $\mathcal{I}$ is called a model *of a TBox $\mathcal{T} = \mathcal{T}_c \cup \mathcal{T}_p$, written $\mathcal{I} \models \mathcal{T}$, if it satisfies each concept inclusion and role inclusion in $\mathcal{T}_c$ and each probabilistic statement in $\mathcal{T}_p$. A TBox $\mathcal{T}$ is* equivalent *to a TBox $\mathcal{T}'$ if $\mathcal{T}$ and $\mathcal{T}'$ have the same models. A TBox $\mathcal{T}$* entails *an axiom α, $\mathcal{T} \models \alpha$, if all models of $\mathcal{T}$ satisfy α. Given probability distribution μ, a TBox $\mathcal{T}$ entails an axiom α under μ, written $\mathcal{T} \models_\mu \alpha$, if all models of $\mathcal{T}$ with probability distribution μ satisfy α.*

[4] A discrete function has a finite or countably infinite set of inputs.

Since μ is a probability distribution, $\sum_{d \in \Delta^{\mathcal{I}}} \mu(d) = 1$. Hence, $\mathbb{P}(\top) = 1$ because $\top$ has the standard DL definition. It also follows from Definition 2 that $\mathbb{P}(\bot) = 0$. Halpern [4] presents an axiom system, which includes standard probabilistic laws, for the Type 1 probabilistic FOL and shows that it is sound. Since the semantics of $\mathcal{SHIQ}_{\mathcal{P}}$ is derived from the Type 1 probabilistic FOL as its fragment, it follows that all standard probabilistic laws hold.

The syntax of $\mathcal{SHIQ}_{\mathcal{P}}$ allows arbitrary linear combinations of probabilities with the same conditioning concept, e.g. considering (2) in Example 1 we could express $\mathbb{P}(FS \mid A) = 1.25 \cdot \mathbb{P}(CS \mid A)$, or “former smokers are 25% more likely than current smokers to be met among adults”.

One may notice that $C \sqsubseteq D$ and $\mathbb{P}(D|C) = 1$ are semantically different. For example, an interpretation $\mathcal{I}$ with $C^{\mathcal{I}} = \emptyset$ is a model of $C \sqsubseteq D$ and not a model of $\mathbb{P}(D|C) = 1$. On the other hand, interpretation $\mathcal{I} = \{\Delta^{\mathcal{I}} = \{a, b\}, C^{\mathcal{I}} = \{a, b\}, D^{\mathcal{I}} = \{b\}, \mu(a) = 0, \mu(b) = 1\}$ is a model of $\mathbb{P}(D|C) = 1$ and not a model of $C \sqsubseteq D$.

An unconditional probability $\mathbb{P}(C)$ is a special case of the conditional probability $\mathbb{P}(C|D)$ with $D \equiv \top$, i.e. $\mathbb{P}(C) = \mathbb{P}(C|\top)$. We can also write all conditional statements in the unconditional form. The following lemma states this.

Lemma 1. *Each probabilistic statement in $\mathcal{SHIQ}_{\mathcal{P}}$ has an equivalent unconditional form $\sum_{j=1}^{m} a_j \mathbb{P}(D_j) \bowtie r$ where $\bowtie \in \{\geq, >\}$, $a_j, r \in \mathbf{R}$, D_j is a $\mathcal{SHIQ}$ concept.*

Proof. See Appendix.

Therefore, we further assume without loss of generality that each statement in $\mathcal{T}_p$ is in unconditional form $\sum_{j=1}^{m} a_j \mathbb{P}(D_j) \bowtie r$, i.e. a linear combination of unconditional probabilities. Thus, a probabilistic TBox $\mathcal{T}_p$ that consists of n probabilistic statements is written as follows:

$$\textstyle\sum_{j=1}^{m_i} a_{ij} \mathbb{P}(D_{ij}) \bowtie r_i, i = 1..n.$$

The *signature* of $\mathcal{T}_c, \mathcal{T}_p$ is the set $\widetilde{\mathcal{T}_c}, \widetilde{\mathcal{T}_p}$ of all concept names and role names occurring in $\mathcal{T}_c, \mathcal{T}_p$, respectively. Thus, $\widetilde{\mathcal{T}} = \widetilde{\mathcal{T}_c} \cup \widetilde{\mathcal{T}_p}$. We call $|\widetilde{\mathcal{T}}|$ the *size* of $\mathcal{T}$. This way of measuring TBox size underestimates the usual size.

3 Illustration and Comparison of the Semantics

Now let us discuss representation capabilities of the statistical and subjective semantics. As mentioned above, the statistical semantics specifies a probability distribution μ over a domain $\Delta^{\mathcal{I}}$, i.e. $P(C^{\mathcal{I}}) = \sum_{d \in C^{\mathcal{I}}} \mu(d)$, whereas the subjective one specifies a probability distribution μ over a set W of possible worlds (which correspond to realizable types), i.e. $P(C) = \sum_{I \in W \mid C \in I} \mu(I)$ [9]. In other words, the statistical semantics represents proportions of domain elements while the subjective one represents degrees of belief. Importantly, in this section the probability distribution μ is fixed, i.e. uniform, and reasoning under restricted distribution may be harder than under unrestricted one due to the reasons discussed in [14]. Let us illustrate some differences between the statistical semantics and the subjective semantics of [9].

Example 2. The following TBox is given:

$$\begin{aligned}\mathcal{T} = \{&H \equiv (= 1\ m.W),\ W \equiv (= 1\ m^{-}.H),\\ &\mathbb{P}(H) = 0.3\},\end{aligned}$$

where H, W are concepts representing husbands and wives, respectively, m is a role representing marriage.

In Example 2, the TBox $\mathcal{T}$ implies that there are *exactly* as many husbands in the domain as there are wives. Hence, given μ is uniform, i.e. all individuals in the domain have the same probability μ_0 to be randomly picked, one can expect $\mathcal{T} \models_\mu \mathbb{P}(W) = 0.3$. However, the subjective semantics is not able to handle this because relationships between individuals within a *single world* are ignored by the semantics since it operates on a set of possible worlds. As a result, the following is entailed: $\mathcal{T} \models \mathbb{P}(W) \geq 0$. On the other hand, the statistical semantics does capture this:

$$\begin{aligned}P(W^{\mathcal{I}}) &= \textstyle\sum_{d \in W^{\mathcal{I}}} \mu(d) = \mu_0 \cdot |W^{\mathcal{I}}|\\ &= \mu_0 \cdot |H^{\mathcal{I}}| = \textstyle\sum_{d \in H^{\mathcal{I}}} \mu(d) = P(H^{\mathcal{I}}).\end{aligned}$$

Example 3 illustrates that there are *at least* as many pets in the domain as there are pet owners. Given μ is uniform, under the statistical semantics the following is entailed: $\mathcal{T} \models_\mu \mathbb{P}(Pe) \geq 0.2$. In contrast, the subjective one gives $\mathcal{T} \models \mathbb{P}(Pe) \geq 0$ due to similar reasons as in Example 2.

Example 3. The following TBox is given:

$$\begin{aligned}\mathcal{T} = \{&PeO \equiv \exists o.Pe,\ Pe \equiv (= 1\ o^{-}.PeO),\\ &\mathbb{P}(PeO) = 0.2\},\end{aligned}$$

where Pe, PeO are concepts representing pets and pet owners, respectively, o is a role representing ownership.

Example 2 and 3 show the capabilities of the statistical semantics to handle statistics. The statistical semantics allows for incorporating prior information about the probability distribution over the domain in a natural way that leads to possibly interesting entailments. In contrast, the subjective semantics ignores relationships between individuals within a single world and is not able to handle proportions.

4 TBox Reasoning in $\mathcal{SHIQ}_\mathcal{P}$

In the context of TBox reasoning in $\mathcal{SHIQ}_\mathcal{P}$ we investigate two reasoning problems: deciding consistency and computing tight probability bounds. We state the problems, develop decision procedures, investigate computational complexity and its sources.

4.1 Deciding Consistency

As usual, a $\mathcal{SHIQ}_\mathcal{P}$ TBox $\mathcal{T} = \mathcal{T}_c \cup \mathcal{T}_p$ is called *consistent* if it admits a model and *inconsistent* otherwise. We call PCon the problem of deciding consistency of a $\mathcal{SHIQ}_\mathcal{P}$ TBox. Further description requires the definition of *types* similar to (complete) types in classic DLs (see e.g. [1]). This should not be confused with the Type 1 and Type 2 logic given by Halpern.

Let $\mathcal{T}$ be a $\mathcal{SHIQ}_\mathcal{P}$ TBox. Let $\mathsf{sub}(\mathcal{T})$ be the set of all subconcepts of concepts occurring in $\mathcal{T}$, $\mathsf{nsub}(\mathcal{T}) = \{\dot\neg C \mid C \in \mathsf{sub}(\mathcal{T})\}$ the set of their negations in negation normal form, i.e. $\dot\neg C = NNF(\neg C)$, and $\mathsf{clos}(\mathcal{T}) = \mathsf{sub}(\mathcal{T}) \cup \mathsf{nsub}(\mathcal{T})$.

Definition 3. *A* type *of $\mathcal{T}$ is a set $t \subseteq \mathsf{clos}(\mathcal{T})$ such that the following conditions are satisfied:*

(i) $C \in t$ iff $\dot\neg C \notin t$, for all $C \in \mathsf{clos}(\mathcal{T})$;
(ii) $C \sqcap D \in t$ iff $C \in t$ and $D \in t$, for all $C \sqcap D \in \mathsf{clos}(\mathcal{T})$;
(iii) $C \sqcup D \in t$ iff $C \in t$ or $D \in t$, for all $C \sqcup D \in \mathsf{clos}(\mathcal{T})$;
(iv) $C \sqsubseteq D \in \mathcal{T}$ and $C \in t$ implies $D \in t$.

Given $\mathcal{I} \models \mathcal{T}$ and $e \in \Delta^\mathcal{I}$, we set $\mathsf{type}(e) = \{D \in \mathsf{clos}(\mathcal{T}) \mid e \in D^\mathcal{I}\}$. We say that a type t of $\mathcal{T}$ is *realized* in a model $\mathcal{I} \models \mathcal{T}$ if there is $e \in \Delta^\mathcal{I}$ such that $\mathsf{type}(e) = t$.

Definition 4. *A type t of $\mathcal{T}$ is* realizable *if $(\sqcap_{C\in t} C)$ is satisfiable w.r.t. $\mathcal{T}$.*

From now on, we consider only realizable types and omit "realizable". It is well-known that satisfiability w.r.t. a $\mathcal{SHIQ}$ TBox is decidable in $\mathsf{ExpTime}$ [19]. Lutz et al. (see Appendix in [15]) state the theorem for complexity of deciding consistency of a Prob1-$\mathcal{ALC}$ TBox and sketch the proof. It can be extended to a $\mathcal{SHIQ}_\mathcal{P}$ TBox.

Theorem 1. *Deciding consistency of a $\mathcal{SHIQ}_\mathcal{P}$ TBox $\mathcal{T}$ is $\mathsf{ExpTime}$-complete.*

Proof. Given a $\mathcal{SHIQ}_\mathcal{P}$ TBox $\mathcal{T} = \mathcal{T}_c \cup \mathcal{T}_p$, we can compute the set T_c of all types of $\mathcal{T}_c$ in $\mathsf{ExpTime}$. It is well-known for DLs with expressivity up to $\mathcal{SHIQ}$ that models are preserved under disjoint union (see e.g. [13]). This implies that there is always a $\mathcal{SHIQ}$ model $(\Delta^\mathcal{I}, \cdot^\mathcal{I}) \models \mathcal{T}_c$ where all (realizable) types are realized. If $\widetilde{\mathcal{T}_p} \backslash \widetilde{\mathcal{T}_c} \neq \emptyset$, T_c is trivially extended to match $\widetilde{\mathcal{T}}$ and denoted T.

Let a variable x_t be associated with each type $t \in T$. Then, by Definition 2 and Lemma 1, the TBox $\mathcal{T}$ induces the system of linear inequalities:

$$E(\mathcal{T}) := \begin{cases} \sum_{t\in T} x_t = 1;\ x_t \geq 0 \text{ for each } t \in T; \\ \sum_{j=1}^{m_i} a_{ij} \sum_{D_{ij}\in t} x_t \bowtie r_i,\ i = 1..n \end{cases}$$

System $E(\mathcal{T})$ can be solved using linear programming with the constant objective. Since linear programming is in P [18] and $E(\mathcal{T})$ is of exponential size in $\mathcal{T}$ we can decide in $\mathsf{ExpTime}$ whether there is a solution. It is sufficient to show that $E(\mathcal{T})$ has a solution iff $\mathcal{T}$ is consistent.

"If". Assume that $E(\mathcal{T})$ has a solution $\dot{X} = \{\dot{x}_t \mid t \in T\}$. There is a classic model $(\Delta^\mathcal{I}, \cdot^\mathcal{I}) \models \mathcal{T}_c$ where all types are realized. Choose any μ satisfying $\sum_{\mathsf{type}(d)=t} \mu(d) =$

$\dot{x}_t, d \in \Delta^{\mathcal{I}}$, e.g. if $\mathsf{type}(d) = t$ then $\mu(d) = \dot{x}_t/(\#\{e \in \Delta^{\mathcal{I}} \mid \mathsf{type}(e) = t\})$. Let $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}}, \mu)$. Then, by Definition 2 of the semantics and Lemma 1

$$\begin{aligned}\textstyle\sum_{D_{ij} \in t} \dot{x}_t &= \textstyle\sum_{D_{ij} \in t} \sum_{\mathsf{type}(d)=t} \mu(d) \\ &= \textstyle\sum_{d \in D_{ij}^{\mathcal{I}}} \mu(d) = P(D_{ij}^{\mathcal{I}}).\end{aligned}$$

This implies that all probabilistic statements in $\mathcal{T}_p$ are satisfied. Hence, $\mathcal{I} \models \mathcal{T}$.

"Only If". Assume $\mathcal{T}$ is consistent, i.e. there is a model $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}}, \mu) \models \mathcal{T}$. Let $\dot{x}_t := \sum_{\mathsf{type}(d)=t} \mu(d)$ for each $t \in T$. By Definition 2 of the semantics and Lemma 1, each probabilistic statement $\sum_{j=1}^{m_i} a_{ij} P(D_{ij}^{\mathcal{I}}) \bowtie r_i,\ i = 1..n$ holds. We observe that

$$\begin{aligned}P(D_{ij}^{\mathcal{I}}) &= \textstyle\sum_{d \in D_{ij}^{\mathcal{I}}} \mu(d) \\ &= \textstyle\sum_{D_{ij} \in t} \sum_{\mathsf{type}(d)=t} \mu(d) = \sum_{D_{ij} \in t} \dot{x}_t.\end{aligned}$$

Therefore, $\dot{X} = \{\dot{x}_t\}$ is a solution of $E(\mathcal{T})$. □

Reduction of a TBox *entailment* $\mathcal{T} \models C \sqsubseteq D$ to consistency is analogous to $\mathcal{SHIQ}$. By Lemma 1, a probabilistic TBox entailment is reduced to consistency as follows:

$$\begin{aligned}&\mathcal{T} \models \textstyle\sum_{j=1}^{m} a_j \mathbb{P}(D_j) \bowtie r \text{ iff} \\ &\mathcal{T} \cup \{(-1) \textstyle\sum_{j=1}^{m} a_j \mathbb{P}(D_j) \ \dot{\bowtie}\ -r\} \text{ is inconsistent,} \\ &\text{where } \dot{\bowtie} = \begin{cases} > \text{ if } \bowtie \text{ is } \geq \\ \geq \text{ if } \bowtie \text{ is } > \end{cases}.\end{aligned}$$

Thus, deciding TBox consistency, and consequently TBox entailment, in $\mathcal{SHIQ}_{\mathcal{P}}$ is not harder than in $\mathcal{SHIQ}$. The procedure is based on solving a system of linear inequalities over variables representing types.

4.2 Computing Tight Probability Bounds

In addition to consistency checking, one may be interested in probabilistic entailments that come in form of *tight bounds* $\mathcal{T} \models \mathbb{P}(C|D) \geq p_\ell$ and $\mathcal{T} \models \mathbb{P}(C|D) \leq p_u$. For example, tight bounds can identify possible data flaws, see Appendix. We define the reasoning task of computing tight probability bounds in $\mathcal{SHIQ}_{\mathcal{P}}$.

Definition 5. *(Tight bounds) Let $\mathcal{T}$ be a consistent $\mathcal{SHIQ}_{\mathcal{P}}$ TBox.*

- *A real value $p_\ell \in [0,\ 1]$ is a* lower bound *for $\mathbb{P}(C|D)$ w.r.t. $\mathcal{T}$ if $\mathcal{T} \models \mathbb{P}(C|D) \geq p_\ell$. p_ℓ is a* tight lower bound *for $\mathbb{P}(C|D)$ w.r.t. $\mathcal{T}$ if p_ℓ is maximal.*
- *A real value $p_u \in [0,\ 1]$ is an* upper bound *for $\mathbb{P}(C|D)$ w.r.t. $\mathcal{T}$ if $\mathcal{T} \models \mathbb{P}(C|D) \leq p_u$. p_u is a* tight upper bound *for $\mathbb{P}(C|D)$ w.r.t. $\mathcal{T}$ if p_u is minimal.*
- *A real value $p_o \in [0,\ 1]$ is called a* tight bound *for $\mathbb{P}(C|D)$ w.r.t. $\mathcal{T}$ if p_o is a tight lower or upper bound.*

- TEnt *is the problem of computing a pair* p_ℓ, p_u *for* $\mathbb{P}(C|D)$*, written* $\mathcal{T} \models_{tight} \{\mathbb{P}(C|D) \geq p_\ell,\ \mathbb{P}(C|D) \leq p_u\}$*. We set* $\ell[\mathbb{P}(C|D)] = p_\ell$ *and* $u[\mathbb{P}(C|D)] = p_u$ *for a tight lower and upper bound of* $\mathbb{P}(C|D)$*, respectively.*

We now state TEnt as an optimization problem.

Lemma 2. *Given a consistent* $\mathcal{SHIQ}_\mathcal{P}$ *TBox* $\mathcal{T}$*, tight bounds* p_ℓ, p_u *for* $\mathbb{P}(C|D)$ *are computed by solving the optimization problem, called* $OP(\mathcal{T})$*:*

$$\begin{array}{ll} \textit{maximise} & s \cdot \dfrac{\sum_{C \sqcap D \in t} x_t}{\sum_{D \in t} x_t} \\ \textit{subject to} & E(\mathcal{T}), \end{array}$$

such that $p_\ell = -p'$, $p_u = p''$*, where* p', p'' *are optimal values of* $OP(\mathcal{T})$ *when* $s = -1$ *and* $s = 1$*, respectively.*

Proof. See Appendix.

As one can see, the objective function in $OP(\mathcal{T})$ is not linear. Fortunately, $OP(\mathcal{T})$ can be translated into an equivalent linear program using a substitution $x_t = y_t/z$ [2]:

$$\begin{array}{ll} \text{maximise} & s \cdot \sum_{C \sqcap D \in t} y_t \\ \text{subject to} & \sum_{t \in T} y_t - z = 0;\ y_t \geq 0 \text{ for each } t \in T; \\ & \sum_{j=1}^{m_i} a_{ij} \sum_{D_{ij} \in t} y_t - r_i z \bowtie 0, i = 1..n; \\ & \sum_{D \in t} y_t = 1;\ z > 0 \end{array}$$

In case of unconditional probability, a substitution is not required (i.e. it is trivial), since the objective function is already linear:

$$\begin{array}{ll} \text{maximise} & s \cdot \sum_{C \in t} x_t \\ \text{subject to} & E(\mathcal{T}) \end{array}$$

Thus, the optimization problem $OP(\mathcal{T})$ is reducible to a linear program. Since linear programming is in P [18], finding a solution of $OP(\mathcal{T})$ requires a polynomial number of iterations in the size of $OP(\mathcal{T})$. Nonetheless, the optimization problem $OP(\mathcal{T})$ is of exponential size w.r.t. the TBox $\mathcal{T}$ due to exponentially many types $t \in T$. Therefore, computing TEnt is in $\mathsf{ExpTime}$ in the size of $\mathcal{T}$ which is stated by the following theorem.

Theorem 2. *Given a consistent* $\mathcal{SHIQ}_\mathcal{P}$ *TBox* $\mathcal{T}$*, computing tight probability bounds is in* $\mathsf{ExpTime}$ *in the size of* $\mathcal{T}$*.*

4.3 More Detailed Complexity Analysis

One can notice that signatures of classic $\widetilde{\mathcal{T}_c}$ and probabilistic part $\widetilde{\mathcal{T}_p}$ do not necessarily coincide. In particular, a probabilistic $\widetilde{\mathcal{T}_p}$ can be much smaller than classic $\widetilde{\mathcal{T}_c}$, $|\widetilde{\mathcal{T}_c}| \gg |\widetilde{\mathcal{T}_p}|$, e.g. for medical knowledge bases. Once realizable types are obtained from $\mathcal{T}_c$,

$\mathcal{T}_p$ produces an optimization problem $OP(\mathcal{T})$ which includes variables for each type from $\mathcal{T}_c$. If $|\widetilde{\mathcal{T}_c}| \gg |\widetilde{\mathcal{T}_p}|$, this makes $OP(\mathcal{T})$ unreasonably large and, consequently, reasoning over relatively simple probabilistic parts hard. Fortunately, as the following lemma shows, this can be avoided via a refinement of $OP(\mathcal{T})$.

Lemma 3. *Given a consistent TBox $\mathcal{T} = \mathcal{T}_c \cup \mathcal{T}_p$, there is a refinement of $OP(\mathcal{T})$, $OP_p(\mathcal{T})$, that gives the same tight bounds as $OP(\mathcal{T})$ and has exponentially many variables in the size of $\mathcal{T}_p$.*

Proof. Let T be the set of all types of $\mathcal{T}$ as above and T_p the set of all types of $\mathcal{T}_p$ alone. Let T' be the set of types of $\mathcal{T}_p$ "allowed" by $\mathcal{T}_c$, i.e. $T' := \{\tau \in T_p \mid \text{there is } t \in T \text{ s.t. } \tau \subseteq t\}$. Then, sums in the optimization problem $OP(\mathcal{T})$ can be rewritten as follows:

$$\sum_{C \in t} x_t = \sum_{C \in \tau} \sum_{\tau \subseteq t} x_t = \sum_{C \in \tau} x_\tau,$$

where $t \in T, \tau \in T'$. Thus, every sum in $OP(\mathcal{T})$, except $\sum_{t \in T} x_t$, is potentially "squeezed" via substitutions, since $|T'| \leq |T|$. Let $T'' = \{t \in T \mid \text{there is no } \tau \in T' \text{ s.t. } \tau \subseteq t\}$. Then $\sum_{t \in T} x_t = \sum_{\tau \in T'} x_\tau + \sum_{t \in T''} x_t = \sum_{\tau \in T'} x_\tau + x$. Constraints $\{x_t \geq 0 \mid \text{there is } \tau \in T' \text{ s.t. } \tau \subseteq t\}$ are substituted by corresponding constraints $\{x_\tau \geq 0\}$. Constraints $\{x_t \geq 0 \mid t \in T''\}$ are substituted by single constraint $\{x \geq 0\}$.

Let $OP_p(\mathcal{T})$ be the optimization problem obtained from $OP(\mathcal{T})$ via the aforementioned substitutions. Since $|T'| \leq |T_p| = 2^{|\widetilde{\mathcal{T}_p}|}$, the number of variables in $OP_p(\mathcal{T})$ is at most $2^{|\widetilde{\mathcal{T}_p}|} + 1$. □

The result of Lemma 3 also holds for deciding TBox consistency in $\mathcal{SHIQ}_\mathcal{P}$.

Corollary 1. *Given a TBox $\mathcal{T} = \mathcal{T}_c \cup \mathcal{T}_p$, there is a refinement of $E(\mathcal{T})$, $E_p(\mathcal{T})$, that gives the same consistency result as $E(\mathcal{T})$ and has exponentially many variables in the size of $\mathcal{T}_p$.*

Lemma 3 gives a procedure to reduce, possibly massively, the number of variables in the linear program for both computing TEnt and deciding PCon. This is achieved via substituting suitable sums of type variables by fresh variables. As a result, the number of variables is reduced from exponentially many w.r.t. $\mathcal{T}$ to exponentially many w.r.t. $\mathcal{T}_p$ only.

It should be noted that Lemma 3 and Corollary 1 do not provide new complexity results: one has to compute the set of realizable types which is still in $\mathsf{ExpTime}$. Nevertheless, it gives insight into complexity sources and may lead to a significant optimization because otherwise the size of a probabilistic part has to be significantly limited, e.g. in [16].

5 Related Work

There are several criteria to distinguish the existing approaches. Firstly, we distinguish *loose* and *tight* proposals to handle probabilities. Loose ones are mainly based on the combination of logic with probabilistic graphical models such as Bayesian networks [10, 20, 3]. They mainly admit a single model. The influence in one direction is typical for them: logical knowledge affects probabilistic knowledge but not the other way around. Among their drawbacks is limited expressivity: the syntax is restricted due to underlying graphical models. In addition, the graphical models often have large sizes that make them hardly manageable. They can also require non-domain assumptions such as probabilistic independences.

By tight proposals we mean those which attempt to deal with uncertainty in purely logical ways. In contrast to loose proposals, they admit multiple models. Influence works in both directions between logical and probabilistic knowledge. The probabilistic DL $\mathcal{SHIQ}_{\mathcal{P}}$ is a tight proposal by this definition.

Probabilistic extensions differ in their syntax and semantics. One of the major syntax differences is how and where probabilities can occur. In particular, probabilistic statements can be added to classical DL axioms [11, 9] or probabilities can be embedded into axioms via application to concepts and roles [15], e.g. $\mathbb{P}_{=0.2}(A) \sqsubseteq CS$ expresses "20 % of adults are current smokers". The syntax of $\mathcal{SHIQ}_{\mathcal{P}}$ is an example of the first approach.

As pointed out above, there are two views of probability that separate the existing probabilistic DLs by their semantics: statistical and subjective. The statistical view has been intensely used by non-logic extensions where a probability distribution is typically represented by graphical models [10, 20, 3]. The examples of logic extensions taking the statistical view are [7, 15]. The subjective view is associated with possible worlds which are, in fact, the core of the semantics for many existing extensions [11, 15, 9]. Lutz et al. [15] obtain the probabilistic DL Prob-$\mathcal{ALC}$ with the subjective semantics from Halpern's Type 2 probabilistic FOL and study its expressivity and computational properties. They also derive Prob1-$\mathcal{ALC}$ from Halpern's Type 1 probabilistic FOL. The statistical semantics of $\mathcal{SHIQ}_{\mathcal{P}}$ is similarly inherited from Halpern's Type 1 probabilistic FOL. A TBox in $\mathcal{SHIQ}_{\mathcal{P}}$ admits linear combinations of conditional probabilities which are not explicitly permitted in Prob1-$\mathcal{ALC}$.

TBox reasoning in $\mathcal{SHIQ}_{\mathcal{P}}$ is the combination of classic DL reasoning and linear programming. The extensions with the subjective semantics [11, 15, 9] also perform reasoning via solving systems of linear inequalities, commonly over possible worlds. This has been implemented and used for medical applications [8].

6 Summary and Future Work

In this work, we study the probabilistic extension of the DL $\mathcal{SHIQ}$ that we call $\mathcal{SHIQ}_{\mathcal{P}}$. It has the statistical semantics inherited from the Type 1 probabilistic FOL [4]. We investigate two reasoning problems for TBoxes in $\mathcal{SHIQ}_{\mathcal{P}}$: deciding consistency PCon and computing tight probability bounds TEnt. We obtain the $\mathsf{ExpTime}$ complexity bounds for deciding consistency of a $\mathcal{SHIQ}_{\mathcal{P}}$ TBox. While deciding consistency of TBoxes is already explored for Prob1-$\mathcal{ALC}$, to the best of our knowledge, no studies are carried out for the problem of computing tight probability bounds for any extension with the same semantics. We state this problem for $\mathcal{SHIQ}_{\mathcal{P}}$ as an optimization problem. We also show that solving this optimization problem is, in fact, reducible to linear programming and, hence, is in P w.r.t. its size. Therefore, computing TEnt is in ExpTime in the size of the TBox $\mathcal{T}$. Thus, both PCon and TEnt are not harder than reasoning in the classic DL $\mathcal{SHIQ}$. Moreover, we show that the size of a linear program for PCon and TEnt can be (significantly) reduced since it depends on the probabilistic part only which is an important insight into the sources of computational complexity.

In Section 3 we discuss abilities of the statistical semantics to handle the statistical knowledge and observe that it is naturally suited for this purpose. As noted by several authors [4, 11, 15], the main shortcoming of the statistical semantics, however, is its inability to represent probabilistic assertional knowledge, i.e. degrees of belief. For example, there is no way to encode "Martin is a smoker with probability 0.7" since he is either a smoker or he is not, according to the statistical view. Lutz et al. [15] state that "only TBox reasoning is relevant" for the statistical semantics and exclude ABoxes.

Nevertheless, we argue that ABox reasoning *may be relevant* for the statistical semantics if it is extended to capture population wide, incomplete data, i.e. it is of sufficient size and spread w.r.t. the whole population. Reasoning over such data might help to answer the question how well the data fits the background knowledge including probabilistic statements, i.e. whether they are compatible with the proportions of individuals in the data. In case of incomplete data, the incompatibility might show whether and which information is missing. Thus, we plan to further study ABox reasoning in the statistical semantics and its possible extensions.

As Example 2 and 3 show, prior knowledge about probability distribution acts as an additional parameter to the reasoning procedure. We plan to further investigate how restrictions of a probability distribution affect entailments and their complexity. We also consider extending expressivity of probabilistic statements in TBoxes, e.g. with probabilistic independence constraints, and investigate related complexity issues. We plan implementations of developed procedures and their optimizations.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. CUP (2003)
2. Charnes, A., Cooper, W.W.: Programming with linear fractional functionals. Naval Research Logistics Quarterly 9(3-4), 181–186 (1962)
3. Costa, P.C.G., Laskey, K.B.: PR-OWL: A framework for probabilistic ontologies. In: Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006). pp. 237–249. IOS Press, Amsterdam, The Netherlands, The Netherlands (2006)
4. Halpern, J.Y.: An analysis of first-order logics of probability. Artificial Intelligence 46, 311–350 (1990)
5. Horrocks, I., Patel-Schneider, P., Van Harmelen, F.: From SHIQ and RDF to OWL: The making of a web ontology language. J. of Web Semantics 1(1), 7–26 (2003)
6. Horrocks, I., Sattler, U., Tobies, S.: Practical reasoning for very expressive description logics. Logic Journal of the IGPL 8, 2000 (2000)
7. Jaeger, M.: Probabilistic role models and the guarded fragment. In: In Proceedings IPMU-2004. pp. 235–242 (2004)
8. Klinov, P., Parsia, B.: Pronto: A practical probabilistic description logic reasoner. In: Bobillo, F., da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW (LNCS Vol.). Lecture Notes in Computer Science, vol. 7123, pp. 59–79. Springer (2013)
9. Klinov, P., Parsia, B., Sattler, U.: On correspondences between probabilistic first-order and description logics. In: Grau, B.C., Horrocks, I., Motik, B., Sattler, U. (eds.) Description Logics. CEUR Workshop Proceedings, vol. 477. CEUR-WS.org (2009)
10. Koller, D., Levy, A., Pfeffer, A.: P-CLASSIC: A tractable probabilistic description logic. In: In Proceedings of AAAI-97. pp. 390–397 (1997)
11. Lukasiewicz, T.: Expressive probabilistic description logics. Artificial Intelligence 172(6-7), 852–883 (2008)
12. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the semantic web. J. Web Sem. 6(4), 291–308 (2008)
13. Lutz, C., Piro, R., Wolter, F.: Description logic tboxes: Model-theoretic characterizations and rewritability. In: Walsh, T. (ed.) IJCAI. pp. 983–988. IJCAI/AAAI (2011)
14. Lutz, C., Sattler, U., Tendera, L.: The complexity of finite model reasoning in description logics. Inf. Comput. 199(1-2), 132–171 (May 2005)
15. Lutz, C., Schröder, L.: Probabilistic description logics for subjective uncertainty. In: Lin, F., Sattler, U., Truszczynski, M. (eds.) KR. AAAI Press (2010)
16. Näth, T.H., Möller, R.: ContraBovemRufum: A system for probabilistic lexicographic entailment. In: Description Logics (2008)
17. Price, C., Spackman, K.: SNOMED clinical terms. British Journal of Healthcare Computing and Information Management 17(3), 27–31 (2000)
18. Schrijver, A.: Theory of Linear and Integer Programming. No. 0471982326, 9780471982326, John Wiley & Sons (1998)
19. Tobies, S.: The complexity of reasoning with cardinality restrictions and nominals in expressive description logics. J. Artif. Intell. Res. (JAIR) 12, 199–217 (2000)
20. Yelland, P.M.: An alternative combination of bayesian networks and description logics. In: Cohn, A.G., Giunchiglia, F., Selman, B. (eds.) KR. pp. 225–234. Morgan Kaufmann (2000)