TBox Reasoning in the Probabilistic Description Logic

\( SHIQ^P \)

Viachaslau Sazonau and Uli Sattler

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

Abstract. One shortcoming of classic Description 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 \( SHIQ^P \) 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 \( SHIQ \). We gain insight into complexity sources.

1 Introduction

Description 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 *SHIQ*, called *SHIQ*$_P$, 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 *SHIQ*$_P$. 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 *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.\textsuperscript{2}

2 Syntax and Semantics of *SHIQ*$_P$

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

**Definition 1.** (TBox syntax) A *SHIQ*$_P$ TBox $\mathcal{T}$ is a set $\mathcal{T}_c \cup \mathcal{T}_p$, where $\mathcal{T}_c$ is a classic *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) $\sum_{j=1}^{m'} a'_j \cdot \mathbb{P}(C_j) \triangleright r'$ (unconditional form);
(ii) $\sum_{j=1}^{m''} a''_j \cdot \mathbb{P}(C_j|D) \bowtie r''$ (conditional form);

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

We use the abbreviation $C \equiv D$ for $\{C \subseteq D, D \subseteq C\}$ 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.** (*SHIQ*$_P$ TBox) According to statistics on smoking in England in 2010,\textsuperscript{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),

\textsuperscript{2}www.cs.man.ac.uk/˜sazonauv/SazonauDL15.pdf
\textsuperscript{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:

\[ \mathcal{T} = \{ S \subseteq A, CS \subseteq S, FS \subseteq S, FS \subseteq \neg CS, S \subseteq CS \cup FS, M \subseteq A, W \subseteq A, M \subseteq \neg W, A \subseteq M \cup W, \]
\[ \mathbb{P}(M | A) = 0.47, \]
\[ \mathbb{P}(CS | A) = 0.2, \mathbb{P}(FS | A) = 0.25, \]
\[ \mathbb{P}(CS | M) = 0.2, \mathbb{P}(FS | M) = 0.29, \]
\[ \mathbb{P}(CS | W) = 0.19, \mathbb{P}(FS | W) = 0.22, \]
\[ \mathbb{P}(CS | \exists m.A) = 0.17, \mathbb{P}(FS | \exists m.A) = 0.28, \]
\[ \mathbb{P}(CS | M \cap \exists m.W) = 0.18, \mathbb{P}(FS | M \cap \exists m.W) = 0.33, \]
\[ \mathbb{P}(FS | W \cap \exists m.M) = 0.17, \mathbb{P}(FS | W \cap \exists m.M) = 0.23, \]
\[ \mathbb{P}(\exists d.(D \cap \exists c.S) | M) = 0.34, \mathbb{P}(\exists d.(D \cap \exists c.S) | W) = 0.25, \]

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 \( \mathcal{SHIQ} \) 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 \( \mathcal{SHIQ} \) TBox is a structure \( \mathcal{I} = (\Delta^\mathcal{I}, \mathcal{I}, \mu) \) where \((\Delta^\mathcal{I}, \mathcal{I})\) is a standard \( \mathcal{SHIQ} \) interpretation and \( \mu \) 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 \( \mathbb{P}(C^\mathcal{I}) \) and

\[ \mathbb{P}(C^\mathcal{I} | D^\mathcal{I}) = \begin{cases} \frac{\mathbb{P}(C^\mathcal{I} \cap D^\mathcal{I})}{\mathbb{P}(D^\mathcal{I})} & \text{if } \mathbb{P}(D^\mathcal{I}) > 0 \\ 0 & \text{otherwise} \end{cases} \]

An interpretation \( \mathcal{I} \) satisfies a probabilistic statement

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

(ii) \( \sum_{j=1}^{m''} a^c_j \cdot \mathbb{P}(C_j | D) \propto r'' \) if \( \sum_{j=1}^{m''} a^c_j \cdot \mathbb{P}(C_j^\mathcal{I} | D^\mathcal{I}) \propto r'' \) 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 \( \alpha \), \( \mathcal{T} \models \alpha \), if all models of \( \mathcal{T} \) satisfy \( \alpha \). Given probability distribution \( \mu \), a TBox \( \mathcal{T} \) entails an axiom \( \alpha \) under \( \mu \), written \( \mathcal{T} \models_\mu \alpha \), if all models of \( \mathcal{T} \) with probability distribution \( \mu \) satisfy \( \alpha \).

\(^4\) A discrete function has a finite or countably infinite set of inputs.
Since $\mu$ is a probability distribution, $\sum_{d \in \Delta^L} \mu(d) = 1$. Hence, $P(\top) = 1$ because $\top$ has the standard DL definition. It also follows from Definition 2 that $P(\bot) = 0$.

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 $SHIQ_P$ is derived from the Type 1 probabilistic FOL as its fragment, it follows that all standard probabilistic laws hold.

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

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

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

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

**Proof.** See Appendix.

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

$$\sum_{j=1}^{m} a_j \sum_{i} P(D_{ij}) \bowtie r, i = 1..n.$$ 

The *signature* of $T_c, T_p$ is the set $\tilde{T}_c, \tilde{T}_p$ of all concept names and role names occurring in $T_c, T_p$, respectively. Thus, $\tilde{T} = \tilde{T}_c \cup \tilde{T}_p$. We call $|\tilde{T}|$ the size of $T$. 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 $\mu$ over a domain $\Delta^L$, i.e. $P(C^I) = \sum_{d \in \Delta^L} \mu(d)$, whereas the subjective one specifies a probability distribution $\mu$ over a set $W$ of possible worlds (which correspond to realizable types), i.e. $P(C) = \sum_{I \in W} \sum_{x \in C} \mu(I)$ [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 $\mu$ 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:

\[ T = \{ H \equiv (= 1 \neg m.W), \ W \equiv (= 1 \neg m.H), \ \mathbb{P}(H) = 0.3 \}, \]

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

In Example 2, the TBox \( T \) implies that there are exactly as many husbands in the domain as there are wives. Hence, given \( \mu \) is uniform, i.e. all individuals in the domain have the same probability \( \mu_0 \) to be randomly picked, one can expect \( T \models \mu \mathbb{P}(W) = 0.3 \). 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: \( T \models \mathbb{P}(W) \geq 0.3 \). On the other hand, the statistical semantics does capture this:

\[
P(W^\mathbb{T}) = \sum_{d \in W^\mathbb{T}} \mu(d) = \mu_0 \cdot |W^\mathbb{T}| = \mu_0 \cdot |H^\mathbb{T}| = \sum_{d \in H^\mathbb{T}} \mu(d) = \mathbb{P}(H^\mathbb{T}).
\]

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

Example 3. The following TBox is given:

\[ T = \{ PeO \equiv \exists o.Pe, \ Pe \equiv (= 1 o^{-}.PeO), \ \mathbb{P}(PeO) = 0.2 \}, \]

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 \( SHIQ_P \)

In the context of TBox reasoning in \( SHIQ_P \) 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 SHIQ 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 SHIQ 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 SHIQ TBox. Let $\text{sub}(\mathcal{T})$ be the set of all subconcepts of concepts occurring in $\mathcal{T}$, $\text{nsub}(\mathcal{T}) = \{\neg C \mid C \in \text{sub}(\mathcal{T})\}$ the set of their negations in negation normal form, i.e. $\neg C = \text{NNF}(\neg C)$, and $\text{clos}(\mathcal{T}) = \text{sub}(\mathcal{T}) \cup \text{nsub}(\mathcal{T})$.

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

(i) $C \in t$ iff $\neg C \notin t$, for all $C \in \text{clos}(\mathcal{T})$;
(ii) $C \cap D \in t$ iff $C \in t$ and $D \in t$, for all $C \cap D \in \text{clos}(\mathcal{T})$;
(iii) $C \cup D \in t$ iff $C \in t$ or $D \in t$, for all $C \cup D \in \text{clos}(\mathcal{T})$;
(iv) $C \subseteq D \in \mathcal{T}$ and $C \in t$ implies $D \in t$.

Given $\mathcal{I} \models \mathcal{T}$ and $e \in \Delta^\mathcal{T}$, we set $\text{type}(e) = \{D \in \text{clos}(\mathcal{T}) \mid e \in D^2\}$. We say that a type $t$ of $\mathcal{T}$ is realized in a model $\mathcal{I} \models \mathcal{T}$ if there is $e \in \Delta^\mathcal{T}$ such that $\text{type}(e) = t$.

Definition 4. A type $t$ of $\mathcal{T}$ is realizable if $(\cap_{C \in t} C)$ 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 SHIQ TBox is decidable in ExpTime [19]. Lutz et al. (see Appendix in [15]) state the theorem for complexity of deciding consistency of a Prob1-ALC TBox and sketch the proof. It can be extended to a SHIQ TBox.

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

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

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

$$E(\mathcal{T}) := \left\{ \begin{array}{l}
\sum_{t \in \mathcal{T}} x_t = 1; \quad x_t \geq 0 \text{ for each } t \in \mathcal{T}; \\
\sum_{j=1}^{n} a_{ij} \sum_{D_j \in t} x_{D_j} \geq r_i, \quad i = 1..n
\end{array} \right\}$$

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 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 $\mathcal{X} = \{x_t \mid t \in \mathcal{T}\}$. There is a classic model $(\Delta^\mathcal{T}_c, \mathcal{T}_c) = \mathcal{T}_c$ where all types are realized. Choose any $\mu$ satisfying $\sum_{\text{type}(d)=t} \mu(d) =$
\( \hat{x}_t, d \in \Delta^2 \), e.g. if \( \text{type}(d) = t \) then \( \mu(d) = \hat{x}_t / (\# \{ e \in \Delta^2 \mid \text{type}(e) = t \}) \). Let \( \mathcal{I} = (\Delta^2, \tau, \mu) \). Then, by Definition 2 of the semantics and Lemma 1
\[
\sum_{D_{ij} \in I} \hat{x}_t = \sum_{D_{ij} \in I} \sum_{\text{type}(d) = t} \mu(d) = \sum_{d \in D_{ij}^2} \mu(d) = P(D_{ij}^2).
\]
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 \( \mathcal{I} = (\Delta^2, \tau, \mu) \models \mathcal{T} \). Let \( \hat{x}_t := \sum_{\text{type}(d) = t} \mu(d) \) for each \( t \in T \). By Definition 2 of the semantics and Lemma 1, each probabilistic statement \( \sum_{j=1}^{m_i} a_{ij} P(D_{ij}^2) \ni r_i, i = 1..n \) holds. We observe that
\[
P(D_{ij}^2) = \sum_{d \in D_{ij}^2} \mu(d) = \sum_{D_{ij} \in I} \sum_{\text{type}(d) = t} \mu(d) = \sum_{D_{ij} \in I} \hat{x}_t.
\]
Therefore, \( \hat{X} = \{ \hat{x}_t \} \) is a solution of \( E(T) \). \( \square \)

Reduction of a TBox entailment \( \mathcal{T} \models C \subseteq D \) to consistency is analogous to \( \mathcal{SHIQ} \). By Lemma 1, a probabilistic TBox entailment is reduced to consistency as follows:
\[
\mathcal{T} \models \sum_{j=1}^{m_i} a_{ij} P(D_{ij}) \ni r \ \text{iff} \ \mathcal{T} \cup \{ (\neg 1) \sum_{j=1}^{m_i} a_{ij} P(D_{ij}) \ni - r \} \text{ is inconsistent,}
\]
where \( \ni \) is minimal.

Thus, deciding TBox consistency, and consequently TBox entailment, in \( \mathcal{SHIQ}_p \) 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 \( \mathcal{T} \models P(C|D) \geq p_L \) and \( \mathcal{T} \models P(C|D) \leq p_U \). For example, tight bounds can identify possible data flaws, see Appendix. We define the reasoning task of computing tight probability bounds in \( \mathcal{SHIQ}_p \).

**Definition 5.** (Tight bounds) Let \( \mathcal{T} \) be a consistent \( \mathcal{SHIQ}_p \) TBox.
- A real value \( p_L \in [0, 1] \) is a lower bound for \( P(C|D) \) w.r.t. \( \mathcal{T} \) if \( \mathcal{T} \models P(C|D) \geq p_L \). \( p_L \) is a tight lower bound for \( P(C|D) \) w.r.t. \( \mathcal{T} \) if \( p_L \) is maximal.
- A real value \( p_U \in [0, 1] \) is an upper bound for \( P(C|D) \) w.r.t. \( \mathcal{T} \) if \( \mathcal{T} \models P(C|D) \leq p_U \). \( p_U \) is a tight upper bound for \( P(C|D) \) w.r.t. \( \mathcal{T} \) if \( p_U \) is minimal.
- A real value \( p_o \in [0, 1] \) is called a tight bound for \( P(C|D) \) 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 \( P(C|D) \), written \( T \models \text{tight} \{ P(C|D) \geq p_l, P(C|D) \leq p_u \} \). We set \( \ell(P(C|D)) = p_l \) and \( u(P(C|D)) = p_u \) for a tight lower and upper bound of \( P(C|D) \), respectively.

We now state TEnt as an optimization problem.

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

\[
\begin{align*}
\text{maximise} & \quad s \cdot \sum_{C \cap D \in t} x_t \\
\text{subject to} & \quad E(T),
\end{align*}
\]

such that \( p_l = -p' \), \( p_u = p'' \), where \( p', p'' \) are optimal values of \( OP(T) \) when \( s = -1 \) and \( s = 1 \), respectively.

**Proof.** See Appendix.

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

\[
\begin{align*}
\text{maximise} & \quad s \cdot \sum_{C \cap D \in t} x_t \\
\text{subject to} & \quad \sum_{t \in T} y_t - z = 0; y_t \geq 0 \text{ for each } t \in T; \\
& \quad \sum_{j=1}^{m_i} a_{ij} \sum_{D \in t} y_t - r_{i} z \geq 0, i = 1..n; \\
& \quad \sum_{D \in t} y_t = 1; z > 0
\end{align*}
\]

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

\[
\begin{align*}
\text{maximise} & \quad s \cdot \sum_{C \in t} x_t \\
\text{subject to} & \quad E(T)
\end{align*}
\]

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

**Theorem 2.** Given a consistent SHIQ P TBox \( T \), computing tight probability bounds is in \( \text{ExpTime} \) in the size of \( T \).

### 4.3 More Detailed Complexity Analysis

One can notice that signatures of classic \( \overline{T}_c \) and probabilistic part \( \overline{T}_p \) do not necessarily coincide. In particular, a probabilistic \( \overline{T}_p \) can be much smaller than classic \( \overline{T}_c \), \( |\overline{T}_c| \gg |\overline{T}_p| \), e.g. for medical knowledge bases. Once realizable types are obtained from \( \overline{T}_c \),
produces an optimization problem \( OP(T) \) which includes variables for each type from \( T_{c} \). If \(|T_{c}| \gg |T_{p}|\), this makes \( OP(T) \) 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 \( OP(T) \).

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

**Proof.** Let \( T \) be the set of all types of \( T \) as above and \( T_{p} \) the set of all types of \( T_{p} \) alone. Let \( T' \) be the set of types of \( T_{p} \) “allowed” by \( T_{c} \), 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 \( OP(T) \) 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 \( OP(T) \), 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 \( \sum_{t \in T} x_{t} = \sum_{t \in T'} x_{t} + \sum_{t \in T''} x_{t} = \sum_{\tau \subseteq t} x_{t} + x \) and \( \sum_{t \in T} x_{t} = \sum_{t \in T'} x_{t} + x \). 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 \( OP_{p}(T) \) be the optimization problem obtained from \( OP(T) \) via the aforementioned substitutions. Since \(|T'| \leq |T_{p}| = 2^{|T_{p}|}\), the number of variables in \( OP_{p}(T) \) is at most \( 2^{|T_{p}|} + 1 \).

The result of Lemma 3 also holds for deciding TBox consistency in \( SHIQP \).

**Corollary 1.** Given a TBox \( T = T_{c} \cup T_{p} \), there is a refinement of \( E(T) \), \( E_{p}(T) \), that gives the same consistency result as \( E(T) \) and has exponentially many variables in the size of \( T_{p} \).

Lemma 3 gives a procedure to reduce, possibly massively, the number of variables in the linear program for both computing \( \text{TEnt} \) and deciding \( \text{PCon} \). 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. \( T \) to exponentially many w.r.t. \( T_{p} \) 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 \( \text{ExpTime} \). 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 $SHIQ_P$ 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. $P_{=0.2}(A) \subseteq CS$ expresses “20% of adults are current smokers”. The syntax of $SHIQ_P$ 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-ALC with the subjective semantics from Halpern’s Type 2 probabilistic FOL and study its expressivity and computational properties. They also derive Prob1-ALC from Halpern’s Type 1 probabilistic FOL. The statistical semantics of $SHIQ_P$ is similarly inherited from Halpern’s Type 1 probabilistic FOL. A TBox in $SHIQ_P$ admits linear combinations of conditional probabilities which are not explicitly permitted in Prob1-ALC.

TBox reasoning in $SHIQ_P$ 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 $SHIQ$ that we call $SHIQ_P$. It has the statistical semantics inherited from the Type 1 probabilistic FOL [4]. We investigate two reasoning problems for TBoxes in $SHIQ_P$: deciding consistency PCon and computing tight probability bounds TEnt. We obtain the ExpTime complexity bounds for deciding consistency of a $SHIQ_P$ TBox. While deciding consistency of TBoxes is already explored for Prob1-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 $SHIQ_P$ 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 $\text{TEnt}$ is in $\text{ExpTime}$ in the size of the $\text{TBox} \ T$. Thus, both $\text{PCon}$ and $\text{TEnt}$ are not harder than reasoning in the classic DL $\text{SHIQ}$. Moreover, we show that the size of a linear program for $\text{PCon}$ and $\text{TEnt}$ 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


7 Appendix

7.1 Examples

Example 4. (Entailment via probability laws) Considering Example 1, let us illustrate how an inference can be derived. Let \( \alpha := P(FS \mid W) = 0.22, T' = T \setminus \alpha \). We wish to compute the exact value of \( P(FS \mid W) \). Since \( W \sqsubseteq A \in T' \), \( T' \models W \sqsubseteq W \sqcap A \). Then, by applying standard probability laws, for any model \( I \models T' \) we derive:

\[
P(FS^I \mid W^I) = \frac{P(FS^I \cap W^I \mid A^I)}{P(W^I \mid A^I)}
\]

\[
= \frac{P(FS^I \mid A^I) - P(FS^I \cap (\neg W)^I \mid A^I)}{1 - P((\neg W)^I \mid A^I)}
\]

\[
= \frac{P(FS^I \mid A^I) - P(FS^I \cap M^I \mid A^I)}{1 - P(M^I \mid A^I)}
\]

\[
= \frac{0.25 - 0.29 \cdot 0.47}{1 - 0.47} = 0.2145
\]

Thus, \( T' \models P(FS \mid W) = 0.2145 \). This entailment contradicts \( \alpha \) because 0.2145 \( \neq \) 0.22. It thereby reveals a flaw, possibly caused by rounding, in the published data.

Example 5. (Computing tight probability bounds) We aim at solving the same problem as in Example 4: to compute the exact value of \( P(FS \mid W) \). However, we do not use probabilistic laws here. Instead, we compute \( \text{TEnt} \):

\[
T \models_{\text{tight}} \{ P(FS \mid W) \geq p_l, P(FS \mid W) \leq p_u \}.
\]

The following TBox \( T' \subset T \), see Example 1, is given:

\[
T' = \{ M \sqsubseteq A, W \sqsubseteq A, M \sqsubseteq \neg W, A \sqsubseteq M \sqcup W, FS \sqsubseteq A, P(M \mid A) = 0.47, P(FS \mid A) = 0.25, P(FS \mid M) = 0.29 \}.
\]

The set of all types for \( T' \) is

\[
T = \{ t_1 := \{ M, \neg W, A, FS \},
      t_2 := \{ \neg M, W, A, FS \},
      t_3 := \{ M, \neg W, A, \neg FS \},
      t_4 := \{ \neg M, W, A, \neg FS \},
      t_5 := \{ \neg M, \neg W, \neg A, \neg FS \} \}.
\]
We introduce a variable \( x_i \) for each \( t_i, i = 1..5 \). Then, the optimization problem \( \text{OP}(\mathcal{T}) \) is stated as follows:

\[
\begin{align*}
\text{maximise} & \quad s \cdot \frac{x_2}{x_2 + x_4} \\
\text{subject to} & \quad x_1 + x_2 + x_3 + x_4 + x_5 = 1; \quad x_i \geq 0, \quad i = 1..5; \\
& \quad x_1 + x_3 - 0.47 \cdot (x_1 + x_2 + x_3 + x_4) = 0; \\
& \quad x_1 + x_2 - 0.25 \cdot (x_1 + x_2 + x_3 + x_4) = 0; \\
& \quad x_1 - 0.29 \cdot (x_1 + x_3) = 0.
\end{align*}
\]

Then, \( \text{OP}(\mathcal{T}) \) is translated into the equivalent linear program using a substitution \( x_i = y_i/z \):

\[
\begin{align*}
\text{maximise} & \quad s \cdot y_2 \\
\text{subject to} & \quad y_1 + y_2 + y_3 + y_4 + y_5 - z = 0; \\
& \quad y_i \geq 0, \quad i = 1..5; \\
& \quad y_1 + y_3 - 0.47 \cdot (y_1 + y_2 + y_3 + y_4) = 0; \\
& \quad y_1 + y_2 - 0.25 \cdot (y_1 + y_2 + y_3 + y_4) = 0; \\
& \quad y_1 - 0.29 \cdot (y_1 + y_3) = 0; \\
& \quad y_2 + y_4 = 1; \quad z > 0.
\end{align*}
\]

If \( s = 1 \), the linear program \( \text{OP}(\mathcal{T}) \) has the optimal solution:

\[
\hat{y} = [0.2572, 0.2145, 0.6296, 0.7855, 75.5858, 77.4726],
\]
where \( \hat{y}_2 = 0.2145 \).

If \( s = -1 \), it has the same optimal solution:

\[
\ddot{y} = [0.2572, 0.2145, 0.6296, 0.7855, 75.5858, 77.4726],
\]
where \( \ddot{y}_2 = 0.2145 \).

Since \( p_u = \hat{y}_2 \) and \( p_v = \ddot{y}_2 \), we obtain:

\[
\mathcal{T} \models_{\text{tight}} \{ \mathbb{P}(FS \mid W) \geq 0.2145, \mathbb{P}(FS \mid W) \leq 0.2145 \}.
\]

Thus, we derive the same entailment as in Example 4:

\[
\mathbb{P}(FS \mid W) = 0.2145.
\]

**Example 6.** (Refinement of an optimization problem) We consider the TBox \( \mathcal{T} \) in Example 5 and add the axiom \( \beta := FS \sqsubseteq S \) to \( \mathcal{T} \). First, we obtain the set of all types for
the resulting TBox $T_5 := T \cup \{ \beta \}$:

$$T = \{ t_1 := \{ M, \neg W, A, FS, S \}, \quad t_2 := \{ \neg M, W, A, FS, S \}, \quad t_3 := \{ M, \neg W, A, \neg FS, S \}, \quad t_4 := \{ M, \neg W, A, \neg FS, \neg S \}, \quad t_5 := \{ \neg M, W, A, \neg FS, S \}, \quad t_6 := \{ \neg M, W, A, \neg FS, \neg S \}, \quad t_7 := \{ \neg M, \neg W, \neg A, \neg FS, S \}, \quad t_8 := \{ \neg M, \neg W, \neg A, \neg FS, \neg S \}. \}$$

As above, we introduce a variable $x_i$ for each $t_i$, $i = 1..8$. Hence, the optimization problem $OP(T)$ now consists of 8 variables instead of 5:

$$\text{maximise} \quad s \cdot \frac{x_2}{x_2 + x_5 + x_6}$$

subject to

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 1;$$

$$x_i \geq 0, \ i = 1..8;$$

$$x_1 + x_3 + x_4 - 0.47 \cdot (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) = 0;$$

$$x_1 + x_2 - 0.25 \cdot (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) = 0;$$

$$x_1 - 0.29 \cdot (x_1 + x_3 + x_4) = 0.$$

The following substitutions take place: $x_{34} := x_3 + x_4, x_{56} := x_5 + x_6, x := x_7 + x_8$. Thus, the total number of variables is reduced from 8 to 5.

### 7.2 Proofs

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

**Proof.** According to Definition 1, each statement in $T_p$ is either in unconditional (i) or conditional (ii) form.

(i) If $\propto \in \{\geq, >\}$, then $m = m', a_j = a'_j, D_j = C_j$ for $j = 1..m', r = r'$. If $\propto \in \{\leq, <\}$, then $m = m', a_j = -a'_j, D_j = C_j$ for $j = 1..m', r = -r'$.

(ii) By Definition 2, $\sum_{j=1}^{m''} a''_j P(C_j | D) \propto r''$ is equivalent to $\sum_{j=1}^{m''} a''_j P(C_j \cap D) - r''. P(D) \propto 0$. Therefore, if $\propto \in \{\geq, >\}$, then $m = m'' + 1, a_j = a''_j, D_j = C_j$ for $j = 1..m'', a_{m''+1} = -r'', D_{m''+1} = D, r = 0.$ If $\propto \in \{\leq, <\}$, then $m = m'' + 1, a_j = a''_j, D_j = C_j$ for $j = 1..m'', a_{m''+1} = r'', D_{m''+1} = D, r = 0.$

\[\Box\]
Lemma 2. Given a consistent $SHIQ_P$ TBox $\mathcal{T}$, tight bounds $p_L, p_u$ for $P(C|D)$ are computed by solving the optimization problem, called $OP(\mathcal{T})$:

$$\text{maximise } \frac{s \cdot \sum_{C\cap D \in \mathcal{T}} x_t}{\sum_{D \in \mathcal{T}} x_t}$$
subject to $E(\mathcal{T})$,

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

Proof. Since $\mathcal{T}$ is consistent, there is an optimal solution. Let this optimal solution be $\hat{X} = \{\hat{x}_t | t \in \mathcal{T}\}$, where $\mathcal{T}$ is the set of all types for $\mathcal{T}$. Let $p_0$ be the value of the objective function with the input $\hat{X}$.

We prove only $p_u = p''$, since $p_L = -p'$ is proved by analogy. Thus, $s = 1$ and $p'' = p_0$. Suppose, the upper bound $p_u \neq p_0$. Then, by Definition 5, there is $p \in [0, 1]$ such that $p > p_0$ and $\mathcal{T}' := \mathcal{T} \cup \{P(C|D) > p_0, P(C|D) \leq p\}$ is consistent. Hence, $E(\mathcal{T}')$ has a solution $\tilde{X} = \{\tilde{x}_t | t \in \mathcal{T}'\}$, where $\mathcal{T}'$ is the set of all types for $\mathcal{T}'$.

If $\tilde{\mathcal{T}}' = \tilde{\mathcal{T}}$, then $\mathcal{T}' = \mathcal{T}$. We can construct a new solution $X = \{x_t | t \in T\}$, where $x_t = \tilde{x}_t$. The value of the objective function given $X$ is $p > p_0$, hence, $\hat{X}$ is not optimal (contradiction). If $\tilde{\mathcal{T}}' \subseteq \tilde{\mathcal{T}}'$, then for each type $\tau \in T'$ there is a type $t \in T$ such that $t \subset \tau$ and for each $t \in T$ there is $\tau \in T'$ such that $t \subset \tau$. We can construct a new solution $X = \{x_t | t \in \mathcal{T}\}$, where $x_t = \sum_{t \subset \tau} \tilde{x}_\tau$. The value of the objective function with $X$ is $p > p_0$, hence, $\hat{X}$ is not optimal (contradiction). $\square$